Critical diplomatic editing

Applying text-critical principles as algorithms

Charles Li[1]

Paper presented at 'Academia, Cultural Heritage, Society' DiXiT Convention, Cologne, March 14-18, 2016.

In recent years, text-based research in the humanities has shifted dramatically from working with critically edited texts to diplomatically-transcribed documents. This is with good reason: both the refinement of computational techniques and the growing interest in the intricacies of textual transmission have led scholars to create archives of transcribed documents in order to facilitate computer-aided textual analysis. But for scholars of an ancient text, for which no autograph exists, it is still vitally important to have a critical edition, with a carefully-curated apparatus, to work with. This is especially evident in the context of Sanskrit texts, some of which exist in dozens, if not hundreds of manuscript witnesses, most of which are extremely corrupt. Many of these documents, when transcribed diplomatically, are simply unreadable.

But nothing prevents us from producing both a critical edition and an archive of document transcriptions that are the source of the edition; in fact, this seems like a natural solution, not only because it facilitates corpus research, but also because it makes the edition much more transparent and open. As a scholarly product, the critical edition should be, in a way, reproducible – the reader should be able to trace an edited passage back to its sources easily, noting precisely what emendations have been made. The critical apparatus traditionally has served as the repository for this information, but, crucially, some silent emendations and omissions – for example, of very common orthographic variants – inevitably are made in order to make the apparatus useful. However, the 'usefulness' of a critical apparatus depends both on the editor's judgment of what to include or exclude and also on a given reader's needs, which may or may not align with the editor's

1 cchli@cantab.net.

critical principles; for example, while the editor may be trying to reconstruct the text as it was composed by the author, the reader may be trying to understand the text as it was read by later commentators. The challenge, then, is to make the critical apparatus flexible – to allow the reader to change the level of detail presented in the apparatus, on demand. Machine collation, applied to diplomatic transcripts, can produce a completely unselective, uncritical apparatus; however, when the collation algorithm is parameterized with a set of critical principles, then a selective apparatus can be generated which can be refined by the reader according to their needs.

The *Dravyasamuddeśa* project

The *Dravyasamuddeśa* project currently is producing an online, digital edition of the *Dravyasamuddeśa* of Bhartṛhari, a Sanskrit text on the philosophy of language, along with the *Prakīrṇaprakāśa* commentary by Helārāja. In order to achieve the aim of realizing an 'open source' edition, each witness is transcribed diplomatically in TEI XML and linked to the edition text.[2] These witnesses then are collated automatically, using the *Myers diff* algorithm (Myers 1986, 251-266), to produce an apparatus. However, since the diplomatic transcripts contain variations in punctuation, orthography, and the application of sandhi rules, the *diff* algorithm naively would report these differences in the apparatus. Therefore, in order to refine the generated apparatus, the web interface of the edition includes a number of options to filter out unwanted information (Figure 1). By using a machine collation algorithm rather than collating manually, the results are more consistent and precise, since a great deal of human error is avoided. Moreover, an apparatus can be generated automatically for any witness as a base text; the critical text no longer has the same privileged status as in print editions, where the witness texts exist only as apparatus variants. All of the diplomatically transcribed witnesses are treated as texts in their own right and are fully searchable. But perhaps most importantly, working with diplomatic transcripts and machine collation forces the editor to express their text-critical principles in a precise and formal manner, as machine-readable algorithms.

Text-critical principles as regular expressions

In order to filter out unwanted entries from the automatically-generated apparatus, the diplomatic transcripts are pre-processed prior to being collated. The pre-processing is performed in three stages: first, XML tags are stripped, along with tagged content that should be ignored (such as marginal notes and deleted text); secondly, punctuation as well as other irrelevant characters, such as digits, are removed; and finally, the orthography is normalized according to a set of text-critical principles. This last operation, normalization, is achieved by expressing the text-critical principles as regular expressions.

---

2 See Formigatti (forthcoming), section 3. 2, for an overview of applying TEI to South Asian manuscript sources.

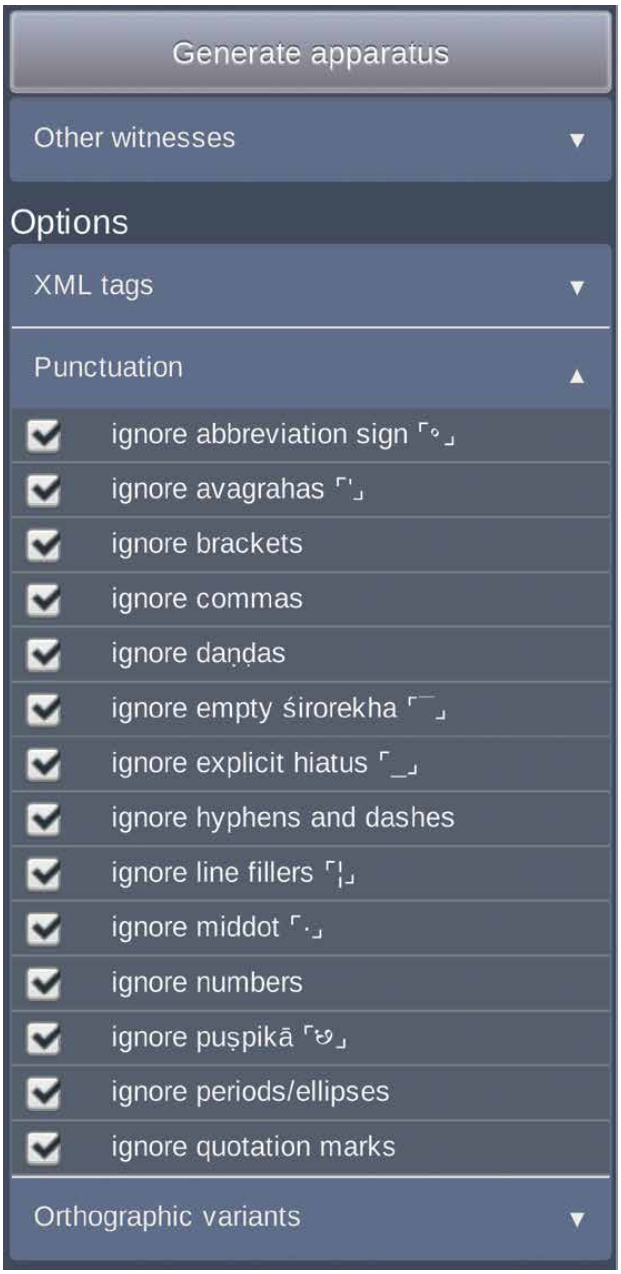

Figure 1: Apparatus options.

Regular expressions are a way of formally describing a search pattern, and they are implemented in most programming languages.[3] For the purposes of normalizing a text for machine collation, regular expressions can be used to replace orthographic variants with their normalized counterparts. In a typical print edition, the text-critical principles that dictate what kinds of variation are ignored are stated in the preface, and those principles are applied silently as the editor collates the witnesses. Even digital projects that use computer software to analyze textual variation usually emend the source texts, rather than work with diplomatic transcriptions; for example, take this recent project at the University of Vienna that aims to produce a critical edition of the *Carakasaṃhitā Vimānasthāna*:

3 See Chapter 9 of *The Open Group Base Specifications, Issue 7*.

> *In the first phase of our still-ongoing editorial work, the 'collation, ' all textual witnesses are compared with the widely known edition of Trikamji, that we chose as our standard version. In the course of this comparison all differences in readings between the manuscripts and the text as edited by Trikamji are noted with very few exception, like, for example, sandhi-variants, variants of punctuation, variants of consonant gemination after 'r, ' variants of homograph and semi-homograph akṣaras*
>
> (Maas 2013, 32).

Just as in a traditional critical edition, the witness texts are collated manually, and some variants are discarded completely. But if we employ machine collation, we can transcribe diplomatically all witness texts, and then use text-critical principles, precisely expressed as regular expressions, to normalize them before collating.

Example: normalizing semi-homograph nasals

One common variation that is ignored in critical apparatuses of Sanskrit texts is that of semi-homograph nasals. In most scripts used to write Sanskrit, the nasals *ṅ, ñ, ṇ,* and *n*, along with *m*, often are written as the anusvāra *ṃ*. In most editions, this variation is discarded silently; typically, the editor would express this rule in a prefatory statement, such as 'variants of semi-homograph nasals are not noted'. But with machine collation, this rule must be expressed in a formal language, and this requirement gives us the opportunity to refine our text-critical principle to be as specific as possible. The replacement of a nasal with *ṃ* occurs only under certain specific conditions, and, based both on Sanskrit grammatical theory and a survey of the manuscripts being collated, a formal rule can be devised which expresses these conditions. In the case of semi-homograph nasals, the text can be normalized using the regular expression

/ṅ(?=[kg])|ñ(?=[cj])|ṇ(?=[ṭḌ])|n(?=[tdn])|m(?=[pb])/ṃ/

Expressed in English, this means:
Replace
ṅ when followed by *k* or *g*,
ñ when followed by *c* or *j*,
ṇ when followed by *ṭ* or *ḍ*,
n when followed by *t*, *d*, or *n*,
and *m* when followed by *p* or *b*,
with *ṃ*.

When this regular expression is applied to the texts before they are compared by the *diff* algorithm, the resulting apparatus will not include semi-homograph nasal variants. This approach is preferable to manual collation in a number of respects: firstly, expressing a text-critical principle in a formal language such as a regular expression forces the editor to be as specific and precise as possible; secondly, the diplomatic transcript of the manuscript, with its own particular orthography, is unaffected by the process; and finally, any one of these rules can be turned

off by the reader, resulting in, for example, semi-homograph nasal variants being included in the automatically generated apparatus. As a result, the apparatus that is produced is both precise and flexible.

Conclusion

In 1973, Martin L. West declared machine collation not worthwhile, criticizing it for producing 'a clumsy and unselective apparatus' (West 1973, 71-72). His criticism is strictly correct, and, in fact, it articulates a general problem in data-driven analysis: datasets, even very large ones, contain inherent biases, and a straightforward analysis would simply reproduce those biases.[4] In order to achieve meaningful results, domain-specific knowledge needs to be applied. In the example of normalizing semi-homograph nasals, domain-specific knowledge was acquired – gleaned from Sanskrit grammar as well as experience working with Sanskrit manuscripts – expressed formally as a regular expression, and used as a pre-processing step to a general-purpose algorithm, *Myers diff*. By applying text-critical principles to the task of machine collation, an apparatus can be generated automatically that is neither clumsy nor unselective, and which is more precise than what could have been achieved manually.

Since West made his statement criticizing machine collation, there has been a shift in scholarly attitudes towards what a critical edition is and what it means for an apparatus to be 'selective'. As West himself admits, editors cannot always be trusted, and the critical apparatus is a way for the reader to check the assertions of the editor. But the apparatus itself is also curated by the editor, and it serves to restrict the reader to a very limited perspective of the textual evidence for the edition. For the scholar of an ancient text, this is not enough; new modes of inquiry demand access to more and more information about the source material. In the *Dravyasamuddeśa* project, we hope to facilitate this by making the edition as 'open source' as possible, without sacrificing the intelligibility of a 'selective' critical apparatus; we merely have expressed our selection criteria – our text-critical principles – as filters that can be turned on or off by the reader. In effect, the apparatus is transformed from a static, authoritative presentation of textual evidence to the site of a negotiation between the textual hypothesis of the editor and the analysis of the reader.

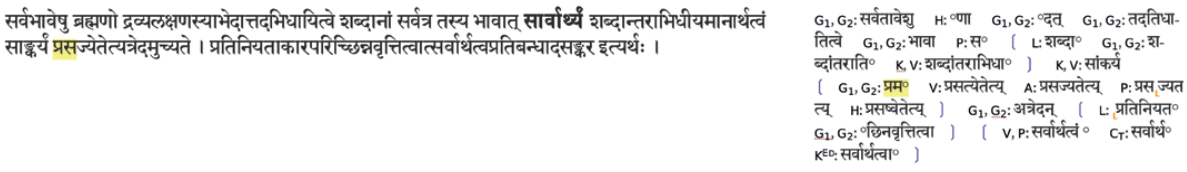

Figure 2: The generated apparatus: clicking on the variant highlights the lemma in the text.

4 For examples from research in the humanities, see Gitelman 2013.

Resources

The software being developed is based on open source libraries and is itself open source; the code is hosted on GitHub:<https://github.com/chchch/upama>. An online demonstration of the edition-in-progress can be found at <http://saktumiva.org/wiki:dravyasamuddesa:start>.

References

Formigatti, Camillo A. Forthcoming. 'From the Shelves to the Web: Cataloging Sanskrit Manuscripts in the Digital Era'. In *Paper & Pixel: Digital Humanities in Indology*. Edited by Elena Mucciarelli and Heike Oberlin. Wiesbaden: Harrassowitz.

Gitelman, Lisa, (ed.) 2013 *'Raw Data' is an Oxymoron*. Cambridge, Mass.: MIT Press.

Institute of Electrical and Electronic Engineers and The Open Group. 2016 'Regular Expressions'. *The Open Group Base Specifications*, Issue 7. Accessed 4 March 2017. http://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap09.html.

Maas, Philipp A. 2013. 'On What to Do with a Stemma – Towards a Critical Edition of the Carakasaṃhitā Vimānasthāna 8. 29'. In *Medical Texts and Manuscripts in Indian Cultural History*, edited by D. Wujastyk, A. Cerulli and K. Preisendanz. New Delhi: Manohar.

Myers, Eugene W. 1986. 'An O(ND) Difference Algorithm and its Variations.' *Algorithmica* 1: 251-266.